Prototype Of Speech Translation System For Audio Effective Communication

Richard Rojas Bello[1], Erick Araya Araya[2] and Luis Vidal Vidal[2]
[1]Escuela de Ingeniería Civil en Informática, Universidad Austral de Chile,
Valdivia, Chile
rrojas1@inf.uach.cl
[2]Instituto de Informática, Universidad Austral de Chile
Valdivia, Chile
{earaya,lvidal}@inf.uach.cl

Abstract. The present document exposes the development of a prototype of translation system as a Thesis Project. It consists basically on the capture of a flow of voice from the emitter, integrating advanced technologies of voice recognition, instantaneous translation and communication over the internet protocol RTP/RTCP (Real time Transport Protocol) to send information in real-time to the receiver. This prototype doesn't transmit image, it only boards the audio stage. Finally, the project besides embracing a problem of personal communications, tries to contribute to the development of activities related with the speech recognition, motivating new investigations and advances on the area.

1 Introduction

At present internet offers almost instantaneous different and efficient ways of communication without considering the distance among people. The current technology allows the access to e-mail, news services, instant messaging services (for example *MSN Messenger*) and applications for video conference. Nevertheless, in the topic of video-conference and specifically in voice conversations there are still obstacles that hinder a full communication; one of them is the difference of languages. This is the point on which the proposed solution is focused on; solution that approaches the problem integrating technologies of recognition and speech synthesis, together with technologies of transmission of voice on IP networks (VoIP) (Rojas Bello 2005).

1.1 State of the art

In contrast to the traditional biometric recognition - as it can be fingerprint - the speech recognition is neither fixed nor static, there is only dependent information of the act.

The state of the art of the automatic verification of the speech proposes the construction of the speaker's stochastic model based on its own characteristics and extracted from carried out trainings.

Bergdata Biometrics GmbH®, for example, differs between high and low level of information for speech recognition (Graevenitz 2001). Inside the high level of information is the dialect, accent, the speech style and the context, all characteristics that at present are recognized only by human beings. The low level of information contemplates rhythm, tone, spectral magnitude, frequencies and bandwidth of the individual's voice; characteristics that are being used in recognition systems. Bergdata finally states that the speech recognition will be complementary to the biometric techniques.

On December 23, 1999, Science Daily made public the information that scientifics at Carnegie Mellon University and its colleagues of C-STAR (Consortium for Speech Translation Advanced Research) would drive an international videoconference. They showed a system web of planning of journeys. The system used translation speech to speech, interpreting six different languages in six different locations around the world. The demonstration was successful; however, they faced problems characteristic of the spontaneous speech: interruptions, doubts and stutterers (ScienceDaily 1999). The software to communicate and to make references to web documents was *JANUS*. *JANUS* has evolved in its own version III. It manages spontaneous spoken dialogues, of conversation, and with a opened up vocabulary in several speech domains (Waibel 2004).

A similar project, where Carnegie Mellon University is also part is *NESPOLE*. *NESPOLE* is a system applied directly to the world of the electronic trade; it allows to the internet users to visit websites and to be connected transparently with a human agent of the company that provides the service (Cattoni y Lazzari 2005).

2 Prototype

2.1 Design

The figure 1 represents the recognition system prototype, translation and transport over RTP proposed as solution (Rojas Bello 2005). The diagram shows the stream sent by a user A (Spanish language) toward a user B (of English language). The user B executes a far prototype's instance.

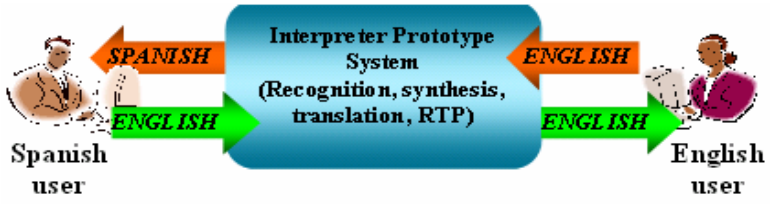

Fig. 1. Interpreter Prototype System

The libraries used to develop the recognition and speech synthesis modules were *Microsoft® SAPI SDK* version 4.0. These libraries add the advantage that the application not only uses the engine included in *SAPI*, but rather engines of other developers can also be recognized by the application if they are compatible with SAPI.

To enable the speech recognition in Spanish the engine *Dragon NaturallySpeaking® Spanish* was used.

The recognition in English was achieved in two different ways, with English Continuous of *Microsoft® SAPI SDK* (included in the libraries) and with *Dragon NaturallySpeaking® English*.

The Spanish synthesis was implemented making use of the *TTS3000 Lernout & Hauspie®* masculine and feminine voices engine; the English synthesis with *Microsoft® SAPI 4.0* voices.

The variety of available voices for speech synthesis offer users different forms of being identified. The prototype has masculine and feminine voices with varied tones and in diverse environments.

The translation was achieved using the service of translation *Google™ Language Tools*, which provides quick translations with appropriate semantics.

Libraries RTP that constitute the base of the audio transmission and reception are *WinRTP* those are open source and are inside present technologies in the market as part of solutions *Cisco AVVID®*.

To view in more detail the figure 2 shows the system's components interaction. In this figure a user speech a phrase which is captured as text using the libraries *SAPI SR* and speech recognitions engines installed previously on the system. Next, the text is send to *Google™ Language Tools* to translate it into Spanish or English according to the case. *Language Tools* returns the text translated and this is synthesized into a WAV file by the libraries *SAPI TTS* and text to speech engines. The WAV files are read by *WinRTP* libraries and sent through internet as RTP packages to the remote user.

The RTP packages that become from the remote user (his/her WAV file synthesized) are received by *WinRTP* and played directly on speakers.

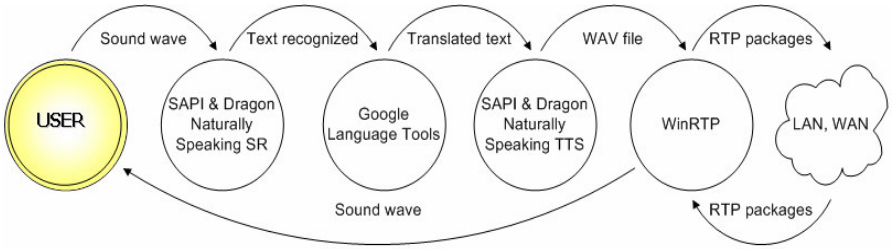

Fig. 2. Interpreter Prototype System's components

2.2 Implementation

The application begins with the dialogue of the figure 3. The user must enter a name and choose the mode of speech recognition that needed.

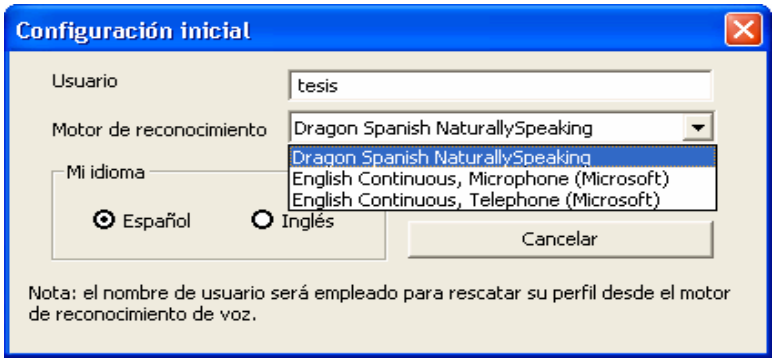

Fig. 3. Users configuration

Next, the user must choose a voice that represents it (fig. 4). The chosen voice will be the one that the second participant of the session will listen. Every time that the user selects a type of voice of the menu the selected character will be introduced itself and told that from that moment it will be the output synthesized voice.

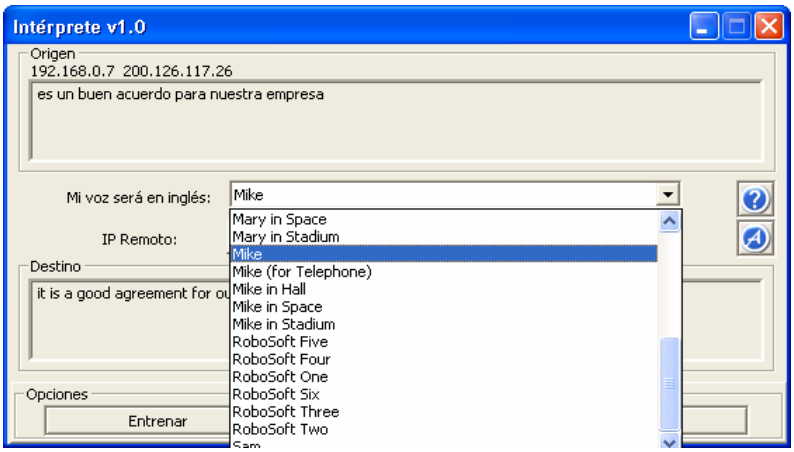

Fig. 4. Main dialogue

For the speaker is necessary to know the IP number of the second computer that executes an application's instance (fig. 5). The communication for this prototype is unicast and begins when pressing the "Transmitir" (Transmit) button. This button has a double purpose: to begin, and to finish the transmission and reception.

Fig. 5. Transmission to remote IP

Every time that the user speaks a sentence the system takes charge of translating it, to synthesize it in the remote participant's language and to send it in real time. Two text boxes show for separate the original text and the translation to synthesize obtained from *Google™ Language Tools.*

2.3 Performance

The best performance obtained by the prototype was using *Dragon NaturallySpeaking® Spanish*. It obtained a high accuracy to 96% just by 5 sessions of training.

The behavior of the bandwidth (BW) was analyzed executing a prototype's instance. The input/output traffic was observed in intervals of 1 second between a 320/128 Kbps and 512/128 Kbps node. On 320/128 Kbps node the behavior is in figure 6.

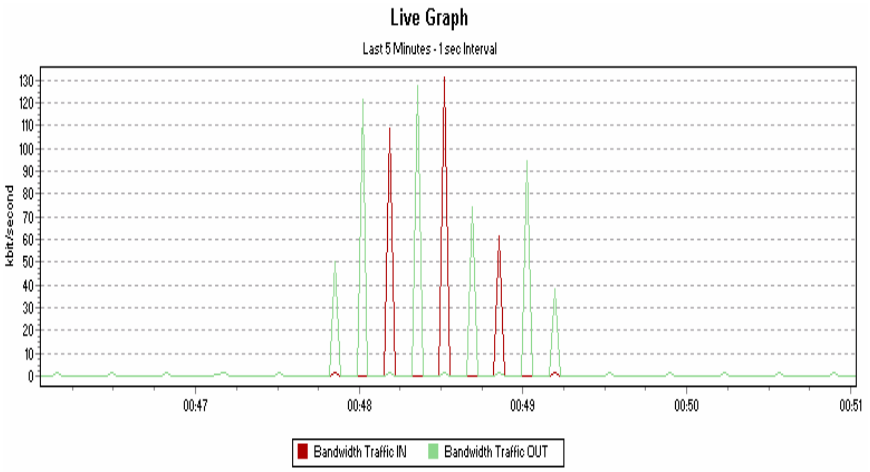

Fig. 6. BW on 320/128 Kbps node

The consumption of BW on the 320/128 Kbps node didn't overcome the 128 Kbps, and the input stream reached the 131 Kbps. The transmitted phrases were received with an approximate retard of 1 sec. and without perceptible jitter.

From the 320/128 Kbps node was carried out the transmission of phrases with different sizes. In the table 1 the sizes are observed (in words and characters) of eleven sentences. Each phrase has associate the necessary time to synthesize it and the maximum consumption of output BW measured in intervals of 1 sec.

Table 1. Performance for characters in 320/128 Kbps connection

Phrase number	Characters	Words	Kbps	Seconds
1	500	77	400	35
2	450	66	400	31
3	400	60	400	27
4	350	50	400	25
5	300	46	385	21
6	250	38	380	18
7	200	32	370	14
8	150	26	350	10
9	100	19	230	7
10	50	9	140	4
11	25	5	75	2

The received audio quality at node 512/256 Kbps was not perceived with jitter until the reception of phrases of 100 characters. The use of BW registered 230 Kbps at node 320/128.

The necessary time to generate the synthesis of the translation behaved in a lineal way (fig. 7), this is, the extension of the phrase results to be proportional at the time used to synthesize it.

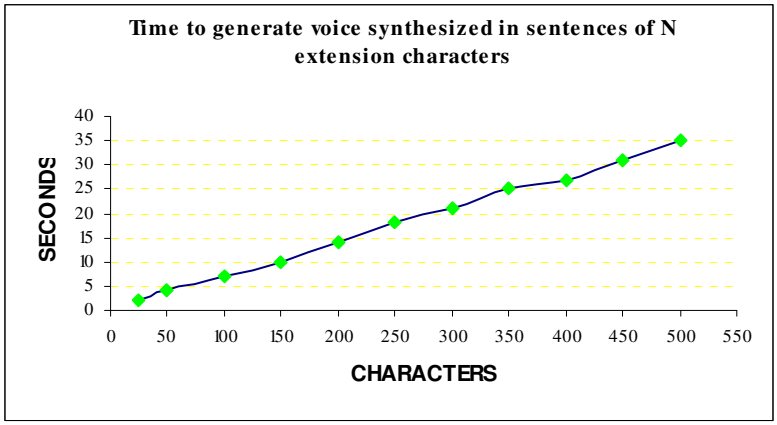

Fig. 7. Time to generate synthesized speech in phrases of N extension characters

3 Conclusions

This project was focused on as an engineering work that seeks for solution to a communicational and social restrictive, the difference of languages. Processes of technological packages opening were faced as and integration of speech recognition, speech synthesis, VoIP and the web service functions.

The designed software can synchronize with the user's needs. The election of compatible libraries with solutions unaware to Microsoft® made the final user to have in his hands the possibility to acquire engines (recognition or synthesis) which are within the user economic reach and adjusted to his specific requirements.

The implemented translation module fulfills the functionality required for the prototype. The connection to *Google™ Language Tools* provided sentences translated successfully in both ways (English/Spanish, Spanish/English) and with a correct syntax and semantics.

The proposed prototype design possesses an expandable connectivity. To incorporate a SIP module to approach the IP and traditional telephony would not represent a radical change in the proposed architecture. Also, SIP uses RTP to transmit information in real time.

The information of the table 1 indicates that the optimum would be to synthesize translations smaller than 50 characters (10 words approx.). An extension of 50 characters doesn't overcome the 4 seconds in synthesizing. The implementation of an algorithm that divides extensive sentences in new of smaller size could reduce even more the time of synthesis.

The use of recognitions engines with high percentages of precision - like *Dragon NaturallySpeaking®* - reduces the occurrence of erroneous recognitions. This directly benefits the efficiency of the recognition module and of the whole system.

References

Rojas Bello R (2005) Diseño y desarrollo de prototipo de sistema de traducción instantánea de habla y transmisión en tiempo real, sobre el protocolo RTP utilizando tecnologías de reconocimiento de voz. Degree Thesis, Universidad Austral de Chile

Graevenitz G (2001). About Speaker Recognition Technology. Bergdata Biometrics GmbH

ScienceDaily (1999). Carnegie Mellon Scientists To Demonstrate Spontaneous Speech-To-Speech Translation In Six Languages. Carnegie Mellon University

Waibel A (2004). Interactive Systems Laboratories. Carnegie Mellon University, Universität Karlsruhe

Cattoni R, Lazzari G (2004). Not only Translation Quality: Evaluating the NESPOLE! Speech-to-Speech Translation System along other Viewpoints. ITC-irst, Carnegie Mellon University, Universität Karlsruhe, CLIPS, University of Trieste, AETHRA